

# CoMBO: Conflict Mitigation via Branched Optimization for Class Incremental Segmentation

Kai Fang<sup>1\*</sup>, Anqi Zhang<sup>1\*</sup>, Guangyu Gao<sup>1†</sup>, Jianbo Jiao<sup>2</sup>, Chi Harold Liu<sup>1</sup>, Yunchao Wei<sup>3</sup>  
<sup>1</sup>Beijing Institute of Technology    <sup>2</sup>University of Birmingham    <sup>3</sup>Beijing Jiaotong University

## Abstract

Effective Class Incremental Segmentation (CIS) requires simultaneously mitigating catastrophic forgetting and ensuring sufficient plasticity to integrate new classes. The inherent conflict above often leads to a back-and-forth, which turns the objective into finding the balance between the performance of previous (old) and incremental (new) classes. To address this conflict, we introduce a novel approach, Conflict Mitigation via Branched Optimization (CoMBO). Within this approach, we present the Query Conflict Reduction module, designed to explicitly refine queries for new classes through lightweight, class-specific adapters. This module provides an additional branch for acquisition of new classes while preserving the original queries for distillation. Moreover, we develop two strategies to further mitigate the conflict following the branched structure, i.e., the Half-Learning Half-Distillation (HDHL) over classification probabilities, and the Importance-Based Knowledge Distillation (IKD) over query features. HDHL selectively engages in learning for classification probabilities of queries that match the ground truth of new classes, while aligning unmatched ones to the corresponding old probabilities, thus ensuring retention of old knowledge while absorbing new classes via learning negative samples. Meanwhile, IKD assesses the importance of queries based on their matching degree to old classes, prioritizing the distillation of important features and allowing less critical features to evolve. Extensive experiments in Class Incremental Panoptic and Semantic Segmentation settings have demonstrated the superior performance of CoMBO. Project Page: <https://guangyu-ryan.github.io/CoMBO>.

## 1. Introduction

Semantic segmentation, the fundamental task in computer vision, involves classifying each pixel into predefined categories. Panoptic segmentation, the challeng-

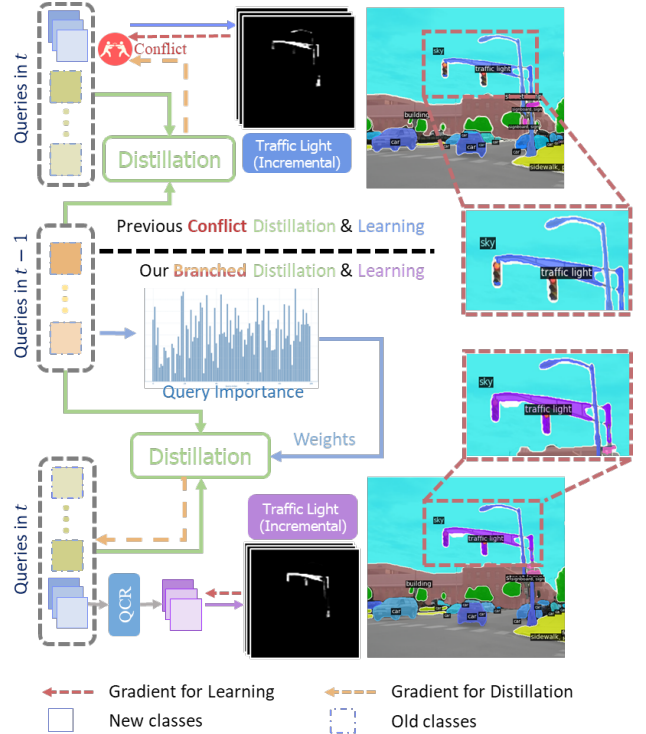


Figure 1. Comparison of our improved distillation and learning strategy (bottom) with previous conflicting strategy (top). Previous strategies impose contradictory supervision on the same target to find a balance, whereas our strategy, including the importance factor for distillation and the QCR module for adaptive learning, enables more compatible, target-specific supervision. The visualization on incremental classes (‘Traffic Light’) highlights the effectiveness of our strategy.

ing variant, unifies semantic and instance segmentation to both classify every pixel and simultaneously identify distinct object instances. Recent advancements in mask classification-based segmentation, such as MaskFormer [16] and Mask2Former [17], enable the unification of semantic and panoptic segmentation. However, traditional approaches to segmentation suffer from limitations in dynamic environments where new classes can emerge unpredictably, as they are typically trained on a static set

\*Equal contribution.

†Corresponding author, guangyugao@bit.edu.cn.

of categories. Therefore, Class Incremental Semantic Segmentation (CISS) has emerged as a pivotal task for continually learning new categories, where models can adapt to new classes without forgetting previously learned ones.

Historically, studies in CISS [3, 35, 49] mainly focuses on tackling *catastrophic forgetting*, employing various techniques such as knowledge distillation [5, 20, 47], freezing parameters [9, 26, 48], or generating pseudo-labels from old model [20]. More recent efforts [8, 11] have extended these techniques to the more challenging Class Incremental Panoptic Segmentation (CIPS) based on the advanced Mask2Former. Nonetheless, these methods often fail to adequately reconcile the inherent tension between retaining old knowledge and acquiring new information. Some overemphasize maintaining stability, thus hindering the learning of new classes, while others prioritize plasticity, leading to significant forgetting of old classes. It is necessary to deconstruct these conflicting objectives to facilitate a more compatible incremental learning process, achieving promotion in both the acquisition of new classes and retention of old classes.

In this work, we propose Conflict Mitigation via Branched Optimization (CoMBO), an approach to reduce conflicts between acquiring new classes and retaining old classes. By integrating three specialized knowledge distillation mechanisms within the Mask2Former architecture, CoMBO effectively mitigates catastrophic forgetting while enhancing the assimilation of new classes. Using bipartite matching, Mask2Former optimally aligns each query, complete with its classification logits and mask predictions, with the most fitting ground truth class, ensuring efficient optimization. Therefore, we first introduce the Half-Learning Half-Distillation mechanism, which selectively applies a shared Kullback-Leibler divergence to the classification logits of unmatched queries. This method distinctly separates distillation on unmatched masks from the classification optimization of matched ones. Moreover, rather than applying uniform distillation across all features, we introduce an importance-based knowledge distillation technique. This technique assesses each query’s significance in preserving knowledge of old classes and adjusts the distillation intensity accordingly, emphasizing the retention of key queries. Additionally, our Query Conflict Reduction module innovatively adapts features for new classes while maintaining the queries’ original characteristics before distillation, using class-specific adapters for each newly learned class. Finally, our approach outperforms previous methods in multiple benchmarks, especially in typical and challenging scenarios such as 100-10 of ADE20K. Our approach achieves remarkable performance of 35.6% and 41.1% mIoU on 100-10 of CIPS and CISS, respectively, compared to the previous state-of-the-art.

Overall, our contributions are summarized as follows:

- We propose a Half-Learning Half-Distillation mechanism that applies soft distillation for old classes and binary-based optimization for new classes to effectively balance retention and acquisition.
- An extra Query Conflict Reduction module is proposed to refine queries for new classes while preserving unrefined features, which enables Importance-based Knowledge Distillation for remembering the key features.
- Through extensive quantitative and qualitative evaluations on the ADE20K dataset, we demonstrate the state-of-the-art performance of our model in both Class Incremental Semantic and Panoptic Segmentation tasks.

## 2. Related Work

### 2.1. Semantic and Panoptic Segmentation

Semantic segmentation aims to classify each pixel in an image, which focuses on the complete coverage of object(s) in each category. The pioneering work FCN [30] and U-Net [36] introduces a pixel-wise classification paradigm by removing the fully connected layers. Since then, numerous approaches [2, 12–15, 29, 38, 41] have been proposed following the paradigm, containing effective modules such as ASPP [13] and pyramid structure [50]. In contrast, panoptic segmentation [27] combines semantic and instance segmentation, requiring the classification of every pixel while distinguishing instances of the same category. These two tasks were addressed separately due to the limitation of the pixel-wise classification. However, the emergence of the mask classification paradigm [16, 17, 43, 45] introduced a universal, transformer-based framework capable of solving multiple segmentation tasks simultaneously. MaskFormer [16] pioneered this paradigm by combining class-agnostic mask proposal generation and mask classification, enabling both semantic and panoptic segmentation simultaneously. Building upon this, Mask2Former [17] enhanced performance by incorporating multi-scale feature fusion and masked attention techniques. However, when learning new categories beyond the initial data, mask classification methods remain susceptible to *catastrophic forgetting*.

### 2.2. Continual Segmentation

In Class Incremental Learning (CIL), fine-tuning models on new data often results in performance drop on old categories, *i.e.*, *catastrophic forgetting* [22, 39]. To address this, various methods [5–7, 20, 21, 32, 33, 51] have been proposed, including retaining representative old samples [1, 4, 9, 10, 31, 32, 35, 53], compensation losses [18, 34, 42, 44], and prompt representations [25, 37, 40]. These advancements in CIL raise interest in more challenging segmentation tasks. Class Incremental Semantic Segmentation (CISS) was first introduced in MiB [5], reconstructing background regions to address *background shifting* during

incremental steps. Since then, most methods [3, 47] have employed techniques such as knowledge distillation [28], pseudo-labels from previous models [8, 20], and background redefinition [5, 46] to mitigate *background shifting*. However, these methods primarily rely on pixel-wise classification models like DeepLabV3 [13]. Therefore, CoMFormer [8] first utilizes the mask classification segmentation model Mask2Former [17] enabling solving both CISS and Class Incremental Panoptic Segmentation (CIPS) tasks. ECLIPSE [26] introduces dynamic model structure expansion methods to extend learnable parameters for new classes and freezes the old parameters. CoMasTRe [23] distills queries matched to old queries with high probability on old classes. BalConpas [11] balance the class proportion on selecting representative replay samples. However, these methods often struggle to balance acquisition and preservation. Our approach mitigates this conflict by separating queries to decouple the two objectives.

### 3. Preliminaries

#### 3.1. Problem Definition

In continual segmentation, the model is tasked with an incremental learning challenge over  $T$  steps. At each step  $t \in \{1, \dots, T\}$ , the model is trained to recognize a unique set of classes  $\mathcal{C}^t$ , where the intersection across all sets from  $i$  to  $T$  is empty  $\bigcap_{t=1}^T \mathcal{C}^t = \emptyset$  and their union forms the complete set of classes  $\bigcup_{t=1}^T \mathcal{C}^t = \mathcal{C}$ . At the current step  $t$ , the training dataset  $\mathcal{D}_{train}^t$  consists of image-label pairs  $(\mathbf{x}^t, \mathbf{y}^t)$ , where  $\mathbf{x}^t$  represents an image and  $\mathbf{y}^t$  its corresponding segmentation label. The labels  $\mathbf{y}^t$  are available only for the classes  $\mathcal{C}^t$  that need to be learned at this step, while labels for previously learned classes  $\mathcal{C}^{1:t-1}$  and future classes  $\mathcal{C}^{t+1:T}$  are inaccessible. After completing training at this step, the model must perform segmentation across all classes learned up to that point,  $\mathcal{C}^{1:t}$ , thereby preventing catastrophic forgetting of the old classes  $\mathcal{C}^{1:t-1}$  while effectively acquiring new ones  $\mathcal{C}^t$ .

#### 3.2. Revisiting Mask2Former and Pseudo-Labeling

The recent advanced universal segmentation network, Mask2Former [17], has favored the class incremental semantic and panoptic segmentation (CISS and CIPS) tasks for its innovative proposal-based structure. Mask2Former deviates from traditional pixel-wise classification pipelines by introducing a class-agnostic mask prediction and classification mechanism. Specifically, the backbone  $\mathbf{f}_b$  and the pixel decoder  $\mathbf{f}_p$  generate multi-scale features that interact with  $N$  learnable queries through multiple self-attention and cross-attention layers in the transformer decoder  $\mathbf{f}_t$ . These  $N$  output queries  $\mathbf{Q} \in \mathbb{R}^{C_Q \times N}$  undergo further processing through two linear layers to form mask embeddings  $\mathcal{E}_{mask} \in \mathbb{R}^{D \times N}$  and class embeddings  $\mathcal{E}_{cls} \in \mathbb{R}^{(C+1) \times N}$ .

The mask embeddings  $\mathcal{E}_{mask}$  yield  $N$  mask predictions  $\mathbf{M} \in \mathbb{R}^{N \times H \times W}$  via dot product between each mask embedding and per-pixel embeddings  $\mathcal{E}_{pixel} \in \mathbb{R}^{D \times H \times W}$  from  $\mathbf{f}_p$ , while  $\mathcal{E}_{cls}$  accounts for the classification predictions of these masks. Note that the additional channel in  $\mathcal{E}_{cls}$  represents the *no-obj* class, an auxiliary category utilized during training but excluded during inference.

In the contexts of CISS and CIPS, where ground truth for old classes  $\mathcal{C}^{1:t-1}$  are inaccessible, pseudo-labeling becomes crucial. PLOP [20] introduces this strategy by generating pixel-level pseudo ground truths for  $\mathcal{C}^{1:t-1}$ , adapted to the mask-based structure by CoMFormer [8], the first method based on Mask2Former for CISS and CIPS tasks. This adaptation involves weighting the mask predictions  $\mathbf{M}$  by the corresponding maximum class embedding score after the softmax process over the  $C + 1$  dimensions, i.e.,  $\max_{c=0}^{|\mathcal{C}^{0:t-1}|} (\text{softmax}_{c=0}^{|\mathcal{C}^{0:t-1}|}(\mathcal{E}_{cls}))$  to form  $\mathbf{M}_\omega$ , with associated categories  $\mathcal{C}_\omega = \arg \max_{c=0}^{|\mathcal{C}^{0:t-1}|} \mathcal{E}_{cls} \in \mathbb{R}^N$ . The pseudo-labeling process then labels the pixels outside the union of  $\mathbf{y}^t$  with the category  $c$  that achieves the maximum score across  $\mathbf{M}_\omega$  and  $\mathcal{C}_\omega$ :

$$\tilde{M}_c(h, w) = \begin{cases} 1 & \text{if } c = \mathcal{C}_\omega[\arg \max_{n=1}^N M_\omega(n, h, w)] \\ & \vee \max_{c \in \mathcal{C}^t} y^t(h, w) = 0, \\ 0 & \text{else} \end{cases} \quad (1)$$

where  $n$  indexes the  $N$  proposals, and  $h, w$  are the pixel coordinates. This pseudo-labeling strategy has become foundational for CISS and CIPS methods [8].

### 4. Proposed Method

In Conflict Mitigation via Branched Optimization (CoMBO), we introduce three distinct strategies: *Half-Distillation-Half-Learning Strategy* for targeted knowledge retention, *Importance-Based Knowledge Distillation* to prioritize crucial features, and *Query Conflict Reduction* for efficient class integration.

#### 4.1. Query Conflict Reduction

As mentioned in Sec. 3.2, both class embeddings  $\mathcal{E}_{cls}$  and mask predictions  $\mathbf{M}$  are generated from the queries  $\mathbf{Q}$ . This interface becomes a critical juncture of conflicts between retaining knowledge of old classes and acquiring new classes manifest, which cannot break through the bottleneck of overall performance by simply finding the balance point.

To deconstruct the crossroad efficiently, we design a Query Conflict Reduction (QCR) module  $f_{QCR}(\cdot)$  as an additional adapter for new classes. As shown in Fig. 2, in most of the scenarios, queries from the previous layer  $\mathbf{Q}_{l-1}$  have the same classification results as the queries from the current layer due to the similar distribution and the same

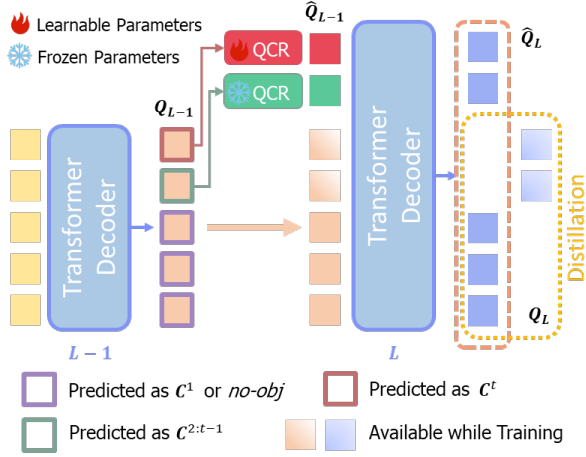


Figure 2. Illustration of the Query Conflict Reduction (QCR) module. This module refines queries that predict incremental classes, allowing the processes of learning new features and retaining old features to occur separately. Note that the QCR module for previous incremental classes is frozen.

classification layer, especially in the latter queries. Due to the shared classifier and mask generator following different layers of the transformer decoder, the classification results on the same position of queries from adjacent layers are approximately the same. Thus, we add the adapter for the queries from the penultimate layer  $Q_{L-1}$  with classification results of new classes  $\mathcal{C}^{2:t}$ , as shown in Fig. 2. To be more specific, the queries  $Q_{L-1, \tilde{c}}$  are selected for the adapter of incremental class  $\tilde{c} \in \mathcal{C}^{2:t}$  via:

$$Q_{L-1, \tilde{c}} = \{Q_{L-1}(n), n \in N \wedge \tilde{c} = \arg \max_{c=0}^{|\mathcal{C}^{0:t}|} \mathcal{E}_{cls}(n)\}, \quad (2)$$

where  $Q_{L-1, \tilde{c}}$  could contain none, one, or some selected queries that enable recognizing different instances. For each new category, we utilize the corresponding QCR module  $f_{QCR, \tilde{c}}(\cdot)$  to refine the group of queries  $Q_{L-1, \tilde{c}}$ . Considering the efficiency and effectiveness, we introduce the low-rank two-layer adaptation as the QCR module:

$$\begin{aligned} \hat{Q}_{L-1, \tilde{c}} &= f_{QCR, \tilde{c}}(Q_{L-1, \tilde{c}}) \\ &= Q_{L-1, \tilde{c}} W_1 W_2 + Q_{L-1, \tilde{c}}, \end{aligned} \quad (3)$$

where  $\hat{Q}_{L-1, \tilde{c}}$  represent the adapted queries of class  $\tilde{c}$ ,  $W_1 \in \mathbb{R}^{D \times r}$  and  $W_2 \in \mathbb{R}^{r \times D}$  denotes the weights for QCR. The QCR module separates the refined queries  $\hat{Q}_{L, \tilde{c}}$  from the original queries  $Q_{L, \tilde{c}}$ , enabling a bifurcate structure for simultaneously learning  $\mathcal{C}^t$  with class-specific adaptation and keep memorizing the features of  $Q_{L, \tilde{c}}$  as well as the logits of  $\mathcal{C}^{0:t-1}$ .

## 4.2. Half-Distillation-Half-Learning Strategy

Previous approaches based on Mask2Former usually separate the learning and distillation process on class embeddings  $\mathcal{E}_{cls}$  due to its different learning strategies compared

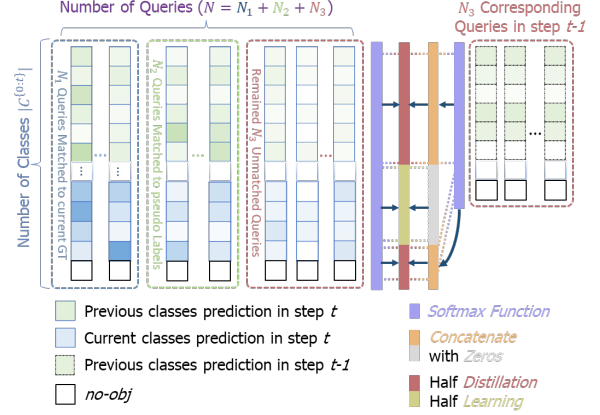


Figure 3. Details of the Half-Distillation-Half-Learning strategy on  $N$  queries, including classification loss on matched queries and Kullback-Leibler Divergence loss on the other queries. The latter involves the logits of both old and current classes.

to pixel-wise segmentation structure (e.g. DeepLabV3), mentioned in Sec. 3.2. This causes conflicts between distillation and learning, where the same logit could have contradictory optimization directions. Therefore, it is necessary to unify the optimization process towards a fixed target.

Our Half-Distillation-Half-Learning (HDHL) strategy combines distillation and learning for  $N$  class embeddings  $\mathcal{E}_{cls}$ , as illustrated in Fig. 3. We follow the bipartite matching of Mask2Former and select embeddings  $\mathcal{E}_{cls}^{\mathcal{C}^t}$  matched to labels of current categories  $\mathcal{C}^t$  and  $\mathcal{E}_{cls}^{\mathcal{C}^{1:t-1}}$  matched to pseudo labels of old categories  $\mathcal{C}^{1:t-1}$ . These embeddings, following the original optimization strategy, are optimized by cross-entropy loss  $\mathcal{L}_{cls}$ . The unmatched embeddings  $\mathcal{E}_{cls}^{\emptyset^t}$  are grouped under the no-object (*no-obj*) category. However, adhering to the original learning strategy leads to *catastrophic forgetting*. Besides, using the distillation strategy for old categories  $\mathcal{C}^{1:t-1}$  using  $\mathcal{E}_{cls}^{\emptyset^{t-1}}$  ignores the learning of current categories  $\mathcal{C}^t$ . However, identifying regions that do not belong to  $\mathcal{C}^t$  is just as important as learning those that do. Considering this, we design a strategy to simultaneously optimize the logits of  $\mathcal{C}^{1:t-1}$  and  $\mathcal{C}^t$  with a shared Kullback-Leibler Divergence  $\mathcal{L}_{kl}$ :

$$\mathcal{L}_{kl} = \frac{1}{|\mathcal{E}_{cls}^{\emptyset^t}|} \sum_{j=1}^{|\mathcal{E}_{cls}^{\emptyset^t}|} \sum_{c=0}^{|\mathcal{C}^{0:t}|} \varphi_{cls}^{\emptyset^{t-1}}(j, c) \log \frac{\varphi_{cls}^{\emptyset^{t-1}}(j, c)}{\varphi_{cls}^{\emptyset^t}(j, c)}, \quad (4)$$

where

$$\begin{aligned} \varphi_{cls}^{\emptyset^t}(j, c) &= \mathcal{S}^t(j, c), \\ \varphi_{cls}^{\emptyset^{t-1}}(j, c) &= \begin{cases} \mathcal{S}^{t-1}(j, c) & \text{if } c \in \mathcal{C}^{0:t-1}, \\ 0 & \text{else,} \end{cases} \end{aligned} \quad (5)$$



and

$$S^t(j, c) = \frac{\exp(\mathcal{E}_{cls}^{\mathcal{O}^t}(j, c))}{\sum_{\hat{c}=0}^{|\mathcal{C}^{0:t}|} \exp(\mathcal{E}_{cls}^{\mathcal{O}^t}(j, \hat{c}))} \quad (6)$$

represent the softmax activation.

The  $\mathcal{L}_{kl}$  not only distill the previous class embedding  $\mathcal{E}_{cls}^{\mathcal{O}^{t-1}}$  to the corresponding categories, including the *no-obj* category, but decreasing the logits of current categories to figure out the objects apart from them. The unified HDHL strategy maintains the probability distribution of old logits seamlessly with the integration of new logits. By combining the original  $\mathcal{L}_{cls}$  and the designed  $\mathcal{L}_{kl}$  as  $\mathcal{L}_{DL}$ , the optimization process covers all the class embeddings without contradictory targets:

$$\mathcal{L}_{DL} = \lambda_{cls} \cdot \mathcal{L}_{cls} + \lambda_{kl} \cdot \mathcal{L}_{kl}, \quad (7)$$

where  $\lambda_{cls}$  and  $\lambda_{kl}$  control the balance between two losses.

### 4.3. Importance-Based Knowledge Distillation

Previous methods mitigate catastrophic forgetting by applying feature distillation, which may hinder the acquisition of new knowledge. To address this, we propose an Importance-Based Knowledge Distillation (IKD) method where stronger distillation is applied to features critical for old classes, while less emphasis on features with less pertinent, thus balancing stability with plasticity.

The importance measurement requires the cost matrix  $\mathcal{A} \in \mathbb{R}^{N \times S}$ , adopted in the bipartite matching process, where  $S$  is the number of image labels. Values in  $\mathcal{A}$  indicate the relationship between queries and classes. For each image, the transformer decoder generates  $N$  predicted segments, denoted as  $\mathcal{Z} = \{(\mathcal{E}_{cls}(n), M(n))\}_{n=1}^N$ . The cost matrix  $\mathcal{A}$  is then computed based on the output segments  $\mathcal{Z}$  and the image labels  $\mathbf{y}^t = \{(c_s, m_s)\}_{s=1}^S$ :

$$\mathcal{A}(n, s) = -\lambda_{cls} \cdot \varphi_{cls}(n, c_s) + \lambda_{mask} \cdot \mathcal{L}_{mask}(\mathcal{E}_{mask}(n), m_s), \quad (8)$$

where  $\varphi_{cls}$  is the classification output obtained by applying softmax to  $\mathcal{E}_{cls}$ , and  $\lambda_{cls}$  and  $\lambda_{mask}$  represent the corresponding hyperparameters.

After each training stage, we compute the cost matrix  $\mathcal{A}$  across every image in the training set  $\mathbf{D}_{train}^t$  of the current incremental step  $t$ . The minimum costs among the classes for  $N$  segments are accumulated in the corresponding position of buffer  $B^t \in \mathbb{R}^N$ , where segments with lower minimum cost values are more likely to recognize and cover the regions of a class  $\mathcal{C}^t$ . By normalizing and reversing the values of  $B^t$ , we derive the importance matrix  $I_{\mathcal{C}^t}^t \in \mathbb{R}^N$  of current classes  $\mathcal{C}^t$ . The importance matrix  $I^{t+1} \in \mathbb{R}^N$  is further estimated via the weighted accumulation of current importance matrix  $I^t \in \mathbb{R}^N$  and  $I_{\mathcal{C}^t}^t$ , where the weights refer to the  $|\mathcal{C}^{1:t-1}|$  and  $|\mathcal{C}^t|$ . The generated  $I^{t+1}$  will be

utilized in the next incremental training phase for weighted distillation of the queries  $Q^t$ , which is computed as follows:

$$\mathcal{L}_{IKD} = \frac{1}{N} \sum_{n=1}^N I^t(n) \cdot \|Q^t(n), Q^{t-1}(n)\|_2^2, \quad (9)$$

where  $\|\cdot\|_2$  denotes the Euclidean distance,  $Q^t(n)$  and  $Q^{t-1}(n)$  denotes the transformer decoder features from the current model  $f^t(\cdot)$  and the old model  $f^{t-1}(\cdot)$ , respectively. Algorithm 1 presents the details of the importance estimation procedure after training  $f^t(\cdot)$ .

---

#### Algorithm 1 Importance Matrix Estimation

---

- 1: **Input:** current model  $f^t(\cdot)$ , current dataset  $\mathbf{D}_{train}^t$ , current importance matrix  $I^t$
  - 2: **Initialization:**  $I^{t+1} \leftarrow 0, B^t \leftarrow 0$
  - 3: **for** all  $(x^t, y^t) \in \mathbf{D}_{train}^t$  **do**
  - 4:   Compute  $\mathcal{A}$  following Eq.(8)
  - 5:   Update the buffer:
  - 6:    $B^t \leftarrow B^t + \min(\mathcal{A}, \text{axis} = 1)$
  - 7: **end for**
  - 8: **for**  $n \in \{1, 2, \dots, N\}$  **do**
  - 9:    $I_{\mathcal{C}^t}^t(n) \leftarrow 1 - \frac{B^t(n) - \min(B^t)}{\max(B^t) - \min(B^t)}$
  - 10:    $I^{t+1}(n) \leftarrow \frac{|\mathcal{C}^{1:t-1}|}{|\mathcal{C}^{1:t}|} \cdot I^t(n) + \frac{|\mathcal{C}^t|}{|\mathcal{C}^{1:t}|} \cdot I_{\mathcal{C}^t}^t(n)$
  - 11: **end for**
  - 12: **Output:** Importance matrix  $I^{t+1}$  for step  $t + 1$
- 

### 4.4. Objective Function

Overall, the objective function of CoMBO is defined as:

$$\mathcal{L}_H = \mathcal{L}_{DL} + \lambda_{IKD} \mathcal{L}_{IKD}. \quad (10)$$

where  $\mathcal{L}_{DL}$  represents the Half-Distillation-Half-Learning loss, and  $\mathcal{L}_{IKD}$  means the Importance-Based Knowledge Distillation loss, with  $\lambda_{IKD}$  as its weighting factor.

## 5. Experimental Results

### 5.1. Experimental Setup

**Datasets and Evaluation Metrics.** We follow the experimental settings of previous works [8, 11, 26] and evaluate our approach on the ADE20K [52] dataset. This dataset is specifically designed to support both panoptic and semantic segmentation tasks. ADE20K contains of 150 classes, including 100 *thing* classes and 50 *stuff* classes. The dataset is composed of 20,210 images for training and 2,000 images for validation.

For Class Incremental Semantic Segmentation (CISS), we adopt the mean Intersection over Union (mIoU) for evaluation. The Intersection over Union (IoU) is calculated as  $\text{IoU} = \frac{TP}{TP + FP + FN}$ , where  $TP$ ,  $FP$ , and  $FN$  denote the

Method	100-50 (2 steps)			100-10 (6 steps)			100-5 (11 steps)			50-50 (3 steps)		
	1-100	101-150	all	1-100	101-150	all	1-100	101-150	all	1-50	51-150	all
FT	0.0	1.3	0.4	0.0	2.9	1.0	0.0	25.8	8.6	0.0	12.0	8.1
MiB [5] <sup>CVPR20</sup>	35.1	19.3	29.8	27.1	10.0	21.4	24.0	6.5	18.1	42.4	15.5	24.4
PLOP [20] <sup>CVPR21</sup>	40.2	22.4	34.3	30.5	17.5	26.1	28.1	15.7	24.0	45.8	18.7	27.7
CoMFormer [8] <sup>CVPR23</sup>	41.1	<b>27.7</b>	36.7	36.0	17.1	29.7	34.4	15.9	28.2	45.0	19.3	27.9
ECLIPSE [26] <sup>CVPR24</sup>	41.7	23.5	35.6	41.4	18.8	33.9	41.1	16.6	<b>32.9</b>	46.0	20.7	29.2
BalCompas [11] <sup>ECCV24</sup>	42.8	<u>25.7</u>	<u>37.1</u>	40.7	<u>22.8</u>	<u>34.7</u>	36.1	<u>20.3</u>	30.8	51.2	<u>26.5</u>	<u>34.7</u>
CoMBO Ours	43.9	25.6	<b>37.8</b>	40.8	<b>25.2</b>	<b>35.6</b>	36.1	<b>20.5</b>	<u>30.9</u>	50.7	<b>28.2</b>	<b>35.7</b>
Joint	43.8	30.9	39.5	43.8	30.9	39.5	43.8	30.9	39.5	50.7	33.9	39.5

Table 1. Quantitative comparison under Class Incremental Panoptic Segmentation with state-of-the-art exemplar-free methods on ADE20K in PQ. Scores of novel classes and all classes in **bold** are the best while underlined are the second best.

number of true-positive, false-positive, and false-negative pixels, respectively. The mIoU metric averages the IoU across all classes for more comprehensive evaluation. For Class Incremental Panoptic Segmentation (CIPS), following previous work [8], we employ Panoptic Quality (PQ) as the evaluation metric. PQ is defined as the product of Recognition Quality (RQ) and Segmentation Quality (SQ). To measure incremental learning capacity, we compute the corresponding metrics for the initial classes  $C^1$ , incremental classes  $C^{2:T}$ , and the aggregate of all classes  $C^{1:T}$ .

**Protocols and Implementation Details.** We follow previous incremental protocols and define scenarios as  $N_{ini} - N_{inc}$ , where  $N_{ini}$  represents the number of initial classes, and  $N_{inc}$  denotes the number of new classes introduced at each incremental step. For example, in the 100-10 scenario, the training begins with 100 classes, followed by the addition of 10 new classes per incremental step, without access to annotations of old classes. For both CIPS and CISS, we do evaluations on the following scenarios: 100-10 (6 steps), 100-50 (2 steps), 100-5 (11 steps) and 50-50 (3 steps).

Our approach is based on the Mask2Former structure [17]. Following previous approaches to CISS and CIPS, we utilize ResNet-50 [24] as backbone for CIPS and ResNet-101 for CISS, with both pre-trained on ImageNet [19]. The input resolution of the images is  $640 \times 640$  with an all-time batch size of 8. We follow the training hyperparameter of Mask2Former in the initial step, with a learning rate of  $10^{-4}$  and iterations of 160,000. During the incremental steps, we set 1,000 iterations per class with a learning rate of  $5 \times 10^{-5}$ . The coefficients  $r$ ,  $\lambda_{cls}$ ,  $\lambda_{kl}$ ,  $\lambda_{IKD}$  are respectively set to 16, 2, 5, 3. All experiments are conducted on the NVIDIA RTX 4090.

## 5.2. Quantitative Results

**Comparisons in Class Incremental Panoptic Segmentation (CIPS).** We evaluated our approach against state-of-the-art exemplar-free methods on the ADE20K dataset

within the CIPS framework, as detailed in Tab. 1. The performance of our CoMBO surpasses other methods in most of the scenarios, particularly in the challenging 100-10 scenario, we achieve 35.6% PQ with at least 3.4% of advantage on incremental classes compared to the previous state-of-the-art methods, demonstrating the effectiveness from reducing the conflicts. Although our approach still underperforms the state-of-the-art approach ECLIPSE on the 100-5 scenario due to its strong memorization ability from freezing parameters, the performance on the new classes remains advanced. Moreover, in the scenario with fewer initial classes, *i.e.* 50-50 scenario, we maintain our advantage in learning new categories with 28.2% PQ for incremental classes and 35.7% PQ for overall performance, widening the gap with the strong distillation methods.

**Comparisons in Class Incremental Semantic Segmentation (CISS).** We further extended our evaluation to the semantic segmentation benchmark, comparing our approach with previous methods on the ADE20K dataset, as detailed in Tab. 2. Our approach consistently outperforms previous methods in all tested scenarios. Notably, in the 100-10 scenario, our approach achieves an mIoU of 41.1%, surpassing the previous state-of-the-art by at least 3.5% for incremental classes, demonstrating its effectiveness in mitigating conflicts. Moreover, in the most challenging 100-5 long-term incremental scenario, our approach not only improves performance on old classes by 2.5% but also achieves a significant 5.4% increase in mIoU for new classes, ensuring both model stability and adaptability.

## 6. Ablation Study

### 6.1. Component Ablations

We analyze the key components of our proposed framework, which includes the Half-Distillation-Half-Learning Strategy, Importance-based Knowledge Distillation, and the Query Conflict Reducing module. Our experiments focus

Method	100-50 (2 steps)			100-10 (6 steps)			100-5 (11 steps)			50-50 (3 steps)		
	1-100	101-150	all	1-100	101-150	all	1-100	101-150	all	1-50	51-150	all
FT	0.0	26.7	8.9	0.0	2.3	0.8	0.0	1.1	0.3	0.0	1.7	1.1
MiB [5] <sup>CVPR20</sup>	37.0	24.1	32.6	23.5	10.6	26.6	21.0	6.1	16.1	45.6	21.0	29.3
PLOP [20] <sup>CVPR21</sup>	44.2	26.2	38.2	34.8	15.9	28.5	39.5	13.6	30.9	54.9	30.2	38.4
CoMFormer [8] <sup>CVPR23</sup>	44.7	26.2	38.4	40.6	15.6	32.3	39.5	13.6	30.9	49.2	26.6	34.1
CoMasTRe [23] <sup>CVPR24</sup>	45.7	26.0	39.2	42.3	18.4	34.4	40.8	15.8	32.6	49.8	26.6	34.5
ECLIPSE [26] <sup>CVPR24</sup>	45.0	21.7	37.1	43.4	17.4	34.6	43.3	16.3	<u>34.2</u>	-	-	-
BalConpas [11] <sup>ECCV24</sup>	49.9	<u>30.1</u>	<u>43.3</u>	47.3	<u>24.2</u>	<u>38.6</u>	42.1	<u>17.2</u>	33.8	55.8	<u>33.3</u>	<u>40.8</u>
CoMBO Ours	50.2	<b>34.4</b>	<b>44.9</b>	47.8	<b>27.7</b>	<b>41.1</b>	44.6	<b>22.6</b>	<b>37.3</b>	55.3	<b>36.9</b>	<b>43.0</b>
Joint	51.7	40.2	47.8	51.7	40.2	47.8	51.7	40.2	47.8	56.6	43.5	47.8

Table 2. Quantitative comparison under Class Incremental Semantic Segmentation with state-of-the-art exemplar-free methods on ADE20K in mIoU. Scores of novel classes and all classes in **bold** are the best while underlined are the second best.

on the 100-10 scenario in the CIPS, as it provides a moderate number of steps to examine the ability of both acquisition and retention. We evaluate various combinations of these components separately and present the results in Tab. 3. The baseline method utilizes the vanilla loss with pseudo-labeling.

HDHL	IKD	QCR	100-10 (6 steps)		
			1-100	101-150	all
			36.3	23.8	32.2
✓			39.4	24.2	34.3
✓	✓		40.0	24.6	34.9
	✓	✓	38.8	25.4	34.4
✓	✓	✓	40.8	25.2	<b>35.6</b>

Table 3. Ablation study of the main components on the 100-10 task of CIPS. The baseline in the 1<sup>st</sup> row employs vanilla losses with pseudo-labeling and excludes the QCR module.

Under identical experimental conditions, the HDHL Strategy significantly enhances the overall PQ by 2.1%, demonstrating its ability to unify logits of new classes into the previous logits distribution without contradictory losses. The implementation of the IKD mechanism has a remarkable 3.7% increase in the PQ for old classes, underscoring its effectiveness in boosting the ability to retain the knowledge of old classes. Furthermore, when combined with the QCR module for branched optimization, CoMBO further boosts both the performance of old and new classes, achieving 0.8% and 0.6% gains in PQ, respectively, surpassing previous methods that struggle to balance acquisition and retention. Consequently, CoMBO achieves a 3.4% PQ improvement over the baseline. The above results demonstrate the effectiveness of the proposed CoMBO and its related modules, showing a noticeable improvement in reducing the conflict on model structures and losses.

## 6.2. Ablation Study of QCR

We validate the effectiveness of QCR with different settings of  $r$  shown in Fig. 5. Under the estimation of performance enhancement and additional parameters, our QCR module achieves the best result when  $r = 16$ , where the quantity of additional parameters comes to 8.2K per class. Considering the total parameter of Mask2Former [17] is 44.9M with ResNet-50, the additional parameters only reach up to 2% of original parameters in total, which is 5 times less than per class additional parameters of ECLIPSE [26] while having better performance on new classes.

## 6.3. Ablation Study of HDHL Strategy

To validate the effectiveness of our Half-Distillation-Half-Learning strategy, we conduct a series of experiments among other strategies, as shown in Tab. 4. The 1<sup>st</sup> row with  $\mathcal{L}_{cls}$  on all classification embeddings represents the vanilla loss with pseudo-labeling, which is adopted in [8]. The 2<sup>nd</sup> row utilizes an auxiliary distillation on the embeddings following [23]. Compared to these strategies with tough supervision on unmatched embeddings  $\mathcal{E}_{cls}^{t,\emptyset}$  and mostly conflicting losses, our HDHL strategy with soft and concentrated supervision has a better PQ of 34.3. Besides, the HDHL design of  $\mathcal{L}_{kl}$  ensures the integrity while optimizing the unmatched embeddings  $\mathcal{E}_{cls}^{t,\emptyset}$ , with improvement of 2.1% compared to distillation-only strategy in the 3<sup>rd</sup> row and 0.8% compared to the simple combination of distillation and  $\mathcal{L}_{cls}$ . The results above show that our HDHL strategy outperforms other learning or distillation strategies on the class embeddings.

## 6.4. Qualitative Analysis

We conduct the qualitative analysis (shown in Fig. 4) by comparing our proposed CoMBO to the *Baseline* (1<sup>st</sup> row of Tab. 3 and *Baseline+HDHL* (2<sup>nd</sup> row of Tab. 3). In the 1<sup>st</sup>, 2<sup>nd</sup>, and 5<sup>th</sup> column, our CoMBO can indicate the incremental classes, *i.e.*, *screen door*, *fountain*, and *ottoman*, while

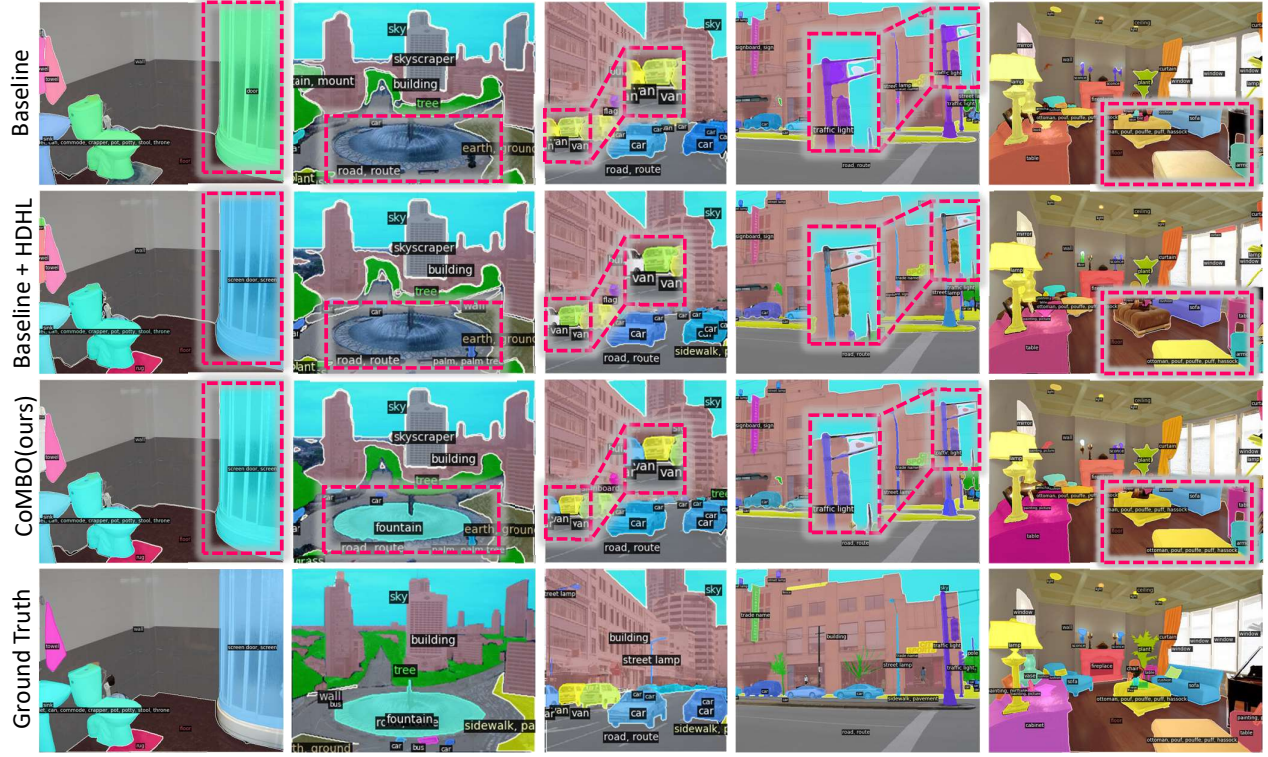


Figure 4. Qualitative results of CoMBO comparing to Baseline, Baseline+HDHL on 100-10 CIPS task of ADE20K.

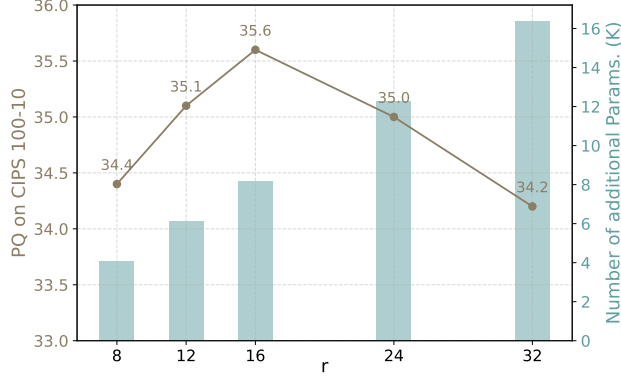


Figure 5. Ablation study of  $r$  in QCR module, including PQ performance (left) and number of parameters (right).

the other methods fail to recognize all of these objects. In the 3<sup>rd</sup> column, our method successfully remembers the *car* in the left instead of covering the object with the prediction of the incremental class *van*. Furthermore, our method allows more precise mask prediction of new class *Traffic Light* in the 4<sup>th</sup> column, demonstrating the effectiveness of refinements on the queries corresponding to new classes.

## 7. Conclusions

In this paper, we present CoMBO, a novel CIS method for mitigating the conflict between acquisition and reten-

$\mathcal{E}_{cls}^{C^t} \& \mathcal{E}_{cls}^{1:t-1}$	$\mathcal{L}_{cls}$	$\mathcal{E}_{cls}^{t,\emptyset}$	$\mathcal{L}_{kl}$	$C^{0:t-1}$	$C^t$	PQ
✓		✓				32.2
✓		✓		✓		32.6
✓			✓			32.2
✓		✓	✓			33.5
✓			✓	✓	✓	<b>34.3</b>

Table 4. Ablation study of the HDHL strategy on task 100-10 of CIPS. The 1<sup>st</sup> row represents the vanilla loss with pseudo-labeling. The last row represents our  $\mathcal{L}_{DL}$ .

tion losses. This conflict arises from the competing goals of efficiently acquiring knowledge about new classes while preserving knowledge of previously learned ones. Our QCR module branches the conflicting optimization targets via lightweight class-specific adaptation on queries, enabling the coexistence of learning and distillation on separated queries. The HDHL strategy and IKD further reduce the contradictory optimization on classification logits and queries with unified targets, respectively, thereby enhancing the overall model performance. Extensive experimental validation on the ADE20K dataset underscores the superiority of our approach, demonstrating the effectiveness of each component in the CIS task, particularly excelling in performance on new classes.



## Acknowledgments

This work was supported by the National Key R&D Program of China (Grant 2022YFC3310200), the National Natural Science Foundation of China (Grant 62472033, 92470203), the Beijing Natural Science Foundation (Grant L242022), the Royal Society grants (SIF\R1\231009, IES\R3\223050) and an Amazon Research Award.

## References

- [1] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Adv. Neural Inf. Process. Syst.*, 32, 2019. 2
- [2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017. 2
- [3] Donghyeon Baek, Youngmin Oh, Sanghoon Lee, Junghyup Lee, and Bumsu Ham. Decomposed knowledge distillation for class-incremental semantic segmentation. *Adv. Neural Inf. Process. Syst.*, 35:10380–10392, 2022. 2, 3
- [4] Zalán Borsos, Mojmir Mutny, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. *Adv. Neural Inf. Process. Syst.*, 33:14879–14890, 2020. 2
- [5] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Buló, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 9233–9242, 2020. 2, 3, 6, 7
- [6] Fabio Cermelli, Dario Fontanel, Antonio Tavera, Marco Ciccone, and Barbara Caputo. Incremental learning in semantic segmentation from image labels. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4371–4381, 2022.
- [7] Fabio Cermelli, Antonino Geraci, Dario Fontanel, and Barbara Caputo. Modeling missing annotations for incremental learning in object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3700–3710, 2022. 2
- [8] Fabio Cermelli, Matthieu Cord, and Arthur Douillard. Comformer: Continual learning in semantic and panoptic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3010–3020, 2023. 2, 3, 5, 6, 7
- [9] Sungmin Cha, YoungJoon Yoo, Taesup Moon, et al. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. *Adv. Neural Inf. Process. Syst.*, 34:10919–10930, 2021. 2
- [10] Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ip, and Sam Kwong. Saving 100x storage: prototype replay for reconstructing training sample distribution in class-incremental semantic segmentation. *Adv. Neural Inf. Process. Syst.*, 36, 2024. 2
- [11] Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ho Shing Ip, and Sam Kwong. Strike a balance in continual panoptic segmentation. In *Proc. Eur. Conf. Comput. Vis.*, pages 126–142. Springer, 2025. 2, 3, 5, 6, 7
- [12] Liang-Chieh Chen. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 2
- [13] L. C. Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint:1706.05587*, 2017. 2, 3
- [14] Liang-Chieh Chen, George Papandreou, and et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017.
- [15] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. Eur. Conf. Comput. Vis.*, pages 801–818, 2018. 2
- [16] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Pixel classification is not all you need for semantic segmentation. *Adv. Neural Inf. Process. Syst.*, 34:17864–17875, 2021. 1, 2
- [17] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1290–1299, 2022. 1, 2, 3, 6, 7
- [18] Wei Cong, Yang Cong, Yuyang Liu, and Gan Sun. Cs2k: Class-specific and class-shared knowledge guidance for incremental semantic segmentation. In *Proc. Eur. Conf. Comput. Vis.*, pages 244–261, 2025. 2
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 248–255, 2009. 6
- [20] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4040–4050, 2021. 2, 3, 6, 7
- [21] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Tackling catastrophic forgetting and background shift in continual semantic segmentation. *arXiv preprint arXiv:2106.15287*, 2021. 2
- [22] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 2
- [23] Yizheng Gong, Siyue Yu, Xiaoyang Wang, and Jimin Xiao. Continual segmentation with disentangled objectness learning and class recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3848–3857, 2024. 3, 7
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, 2016. 6
- [25] M. G. Z. A. Khan, M. F. Naeem, Luc Van Gool, Didier Stricker, Federico Tombari, and Muhammad Zeshan Afzal. Introducing language guidance in prompt-based continual learning. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 11463–11473, 2023. 2
- [26] Beomyoung Kim, Joonsang Yu, and Sung Ju Hwang. Eclipse: Efficient continual learning in panoptic segmentation with visual prompt tuning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3346–3356, 2024. 2, 3, 5, 6, 7
- [27] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proc.*

- IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 9404–9413, 2019. 2
- [28] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 40(12):2935–2947, 2017. 3
- [29] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1925–1934, 2017. 2
- [30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3431–3440, 2015. 2
- [31] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. *Adv. Neural Inf. Process. Syst.*, 30, 2017. 2
- [32] Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. Recall: Replay-based continual learning in semantic segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 7026–7035, 2021. 2
- [33] Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1114–1124, 2021. 2
- [34] Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1114–1124, 2021. 2
- [35] Youngmin Oh, Donghyeon Baek, and Bumsub Ham. Alife: Adaptive logit regularizer and feature replay for incremental semantic segmentation. *Adv. Neural Inf. Process. Syst.*, 35: 14516–14528, 2022. 2
- [36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 2
- [37] Chao Shang, Hongliang Li, Fanman Meng, Qingbo Wu, Heqian Qiu, and Lanxiao Wang. Incrementer: Transformer for class-incremental semantic segmentation with knowledge distillation focusing on old class. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7214–7224, 2023. 2
- [38] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 7262–7272, 2021. 2
- [39] Sebastian Thrun. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. 1998. 2
- [40] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 139–149, 2022. 2
- [41] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1568–1576, 2017. 2
- [42] Jia-Wen Xiao, Chang-Bin Zhang, Jiekang Feng, Xialei Liu, Joost van de Weijer, and Ming-Ming Cheng. Endpoints weight fusion for class incremental semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7204–7213, 2023. 2
- [43] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3014–3023, 2021. 2
- [44] Chaohui Yu, Qiang Zhou, Jingliang Li, Jianlong Yuan, Zhibin Wang, and Fan Wang. Foundation model drives weakly incremental learning for semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 23685–23694, 2023. 2
- [45] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *Proc. Eur. Conf. Comput. Vis.*, pages 288–307, 2022. 2
- [46] Anqi Zhang and Guangyu Gao. Background adaptation with residual modeling for exemplar-free class-incremental semantic segmentation. In *Proc. Eur. Conf. Comput. Vis.*, 2024. 3
- [47] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7053–7064, 2022. 2, 3
- [48] Zekang Zhang, Guangyu Gao, Zhiyuan Fang, Jianbo Jiao, and Yunchao Wei. Mining unseen classes via regional objectness: A simple baseline for incremental segmentation. *Adv. Neural Inf. Process. Syst.*, 35:24340–24353, 2022. 2
- [49] Zekang Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Coinseg: Contrast inter-and intra-class representations for incremental segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 843–853, 2023. 2
- [50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2881–2890, 2017. 2
- [51] Hanbin Zhao, Fengyu Yang, Xinghe Fu, and Xi Li. Rbc: Rectifying the biased context in continual semantic segmentation. In *Proc. Eur. Conf. Comput. Vis.*, pages 55–72, 2022. 2
- [52] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 633–641, 2017. 5
- [53] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Continual semantic segmentation with automatic memory sample selection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3082–3092, 2023. 2